面向 SKA-1 时代的科学数据流及阵列模拟分析

郭绍光[1*]，陆扬[1]，安涛[1]，劳保强[1]，徐志骏[1]，伍筱聪[1]，吕唯佳[1]

1. 中国科学院上海天文台 SKA 区域中心联合实验室, 上海 200030

2. 鹏城实验室 SKA 区域中心联合实验室, 深圳 518066

* 联系人, E-mail: sgguo@shao.ac.cn

收稿日期: 2022-06-28; 接受日期: 2022-07-xx;

SKA 专项 (编号:2020SKA0110300)，国家重点研发计划 (编号:2018YFA0404603)、国家自然科学基金 (批准号:11873079,12041301) 和中国科学院青年创新促进会项目 (编号:2021258) 资助项目

摘要 作为下一代射电望远镜，平方公里阵列望远镜 (SKA) 经过多年的筹备，第一阶段 (SKA1) 已经在 2021 年 7 月开工建设，SKA1 正式运行后预计每年将产生 750PB 的科学归档数据，这些数据将存储在世界各地的 SKA 区域中心供科研工作者使用。本文将SKA观测台站、中央信号处理器、科学数据处理及区域中心等各个阶段的模型进行量化分析，以 SKA1 的高优先级科学观测为主要依据，得出每个阶段的数据流评估情况，以及对科学数据处理算力的需求。以当前SKA1-low 和 SKA1-mid 的阵列为例，总结了包括分辨率、灵敏度、UV 覆盖等影响干涉阵列布局的关键因素；最后使用 OSKAR 进行干涉阵列的数据模拟，通过对SKA1-mid的模拟得出系统的可扩展性和稳定性，通过对SKA1-low在CSRC-P上的模拟，可以看出中国 SKA 区域中心原型机设计经过了充分的论证和优化，并得出了详细的算力需求以及数据量的详细信息。SKA 对数据处理、计算、存储等的需求，将需要电子、通信、信息、计算机等技术和交叉学科的联合推动。

关键词 平方公里阵列射电望远镜, 数据模拟, 综合孔径, 数据格式

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

1 引言

平方公里阵列望远镜（Square Kilometre Array, 简称 SKA ）[1–4] 是一项国际大科学工程，将建造成为世界上最大、最灵敏的射电望远镜。经过多年的准备工作，具备 SKA 望远镜接收能力 10% 的第一阶段SKA-1（SKA Phase 1, 简称 SKA-1）已经于 2021 年 7 月开始建设，SKA-1由位于澳大利亚的SKA1-low和南非的SKA1-mid两个台址组成，覆盖频率为 50MHz ~ 15.3 GHz。

SKA 使用综合孔径技术，将阵列中的信号进行合成处理，提供一个等效口径为 1 公里的射电望远镜。SKA 将产生海量的数据，以SKA1-mid为例，位于阵列附近的中央信号处理器（Centre Signal Process，简称 CSP）对 18Tb/s 的数字化数据流进行波束和相关处理[5]，然后将其传输到开普敦的专用超级计算机进行进一步处理。这些数据流的处理过程对 SKA 整个系统的设计至关重要，也带来了对后续数据管理及数据处理的巨大挑战。根据当前的数据估算以及与全球大型强子对撞机计算网

格（Worldwide Large Hadron Collider Computing Grid，简称 WLCS）的比较，将数据从台址国的科学数据处理器（Science Data Process，简称 SDP）传输到各个 SKA 区域中心（SKA Regional Centre，简称 SRC）需要高速稳定的洲际网络支持[6]。根据每个 SRC 数据分担的评估，为了满足 SKA1 的数据传输，网络带宽至少需要达到 100Gbps 的带宽[7]，才能保证 SKA1 产生的数据稳定及时地到达各个 SRC，SKA1 正式运行后分发到 SRC 的数据流预估为每年 710PB，算力至少需要 25PFlops[1,6]。这给 SKA 设计人员带来了两大技术挑战：如何以较低功耗提供足够的算力来处理数据流；如何将如此大量的数据分发传输到位于全球各地的 SRC[7–9]。

截至到 2021 年底公布的最快的高性能计算机（High Performance Computing，简称 HPC）为安装在日本神户 Riken 计算科学中心的 Fugaku[1)]，HPL 基准测试能力为 442PFlops，该超算基于富士通定制的 ARM A64FX 处理器，并使用富士通的 Tofu D 互联技术进行节点之间的数据传输；另外随着网络新技术的发展，包括多芯光纤技术（Multi-Core Fiber ，简称 MCF）等，当前全球互联网的总带宽已经达到了 618Tbps。与业界最快的 HPC 和全球的互联网络相比较，SKA1 的算力需求与数据传输规模不算很大，但在单一科学研究领域已经属于相当海量的规模，另外由于 SKA 科学研究方向较多，观测模式多样[10–12]，由此导致科学处理的数据流程十分复杂[13]，HPC 通常并不能满足其多样化、定制化的需求，SRC 的处理将需要异构硬件对不同的科学观测使用不同的处理管线[14–16]，而这些管线根据数据密集型、计算密集型、存储密集型等进行分类处理。

本文从阵列的数据接收开始，到最后到达 SRC 的科学数据（含图像、元数据、脉冲星候选体等数据[17]）的详细情况，通过对其主要包括的数据传输速率以及算力需求来介绍 SKA1 的具体数据流，由于 SKA1 的观测配置目前尚未最终固化，当前对数据流仅进行模型预估处理；通过对 SKA1 的阵列布局具体情况及阵列部署的具体原则，对数据流进行定量化的描述；并通过对不同阵列在目前CSRC-P上进行的实验和测试，得出具体数据流，以此来阐述 SKA1 数据流的概况与全规模 SKA 正式运行后，对CSRC-P计算、存储和算力的影响。从而给出对未来CSRC-P数据模拟的一些建议和规划。

1) https://www.top500.org/

2 SKA 数据流

来自射电源的无线电信号到达 SKA 望远镜阵列后被天线接收，经过模拟数字转换器（Analog to digital converter，简称 ADC）转换为数字信号，随后数据流进入中央信号处理系统（Center Signal Processing，简称 CSP）进行相关处理，输出可见度数据，这些数据将经过初步预校准，产生校准数据、图像数据和其他元数据，经深度处理后，这些数据将开放给科学团队，进行后续的科学研究工作。

如图1所示，在整个数据观测处理流程中，SKA1-low 的外围站通过远程处理设施（Remote Processing Facilities，简称 RPF）处理后与核心站通过中央处理设备（Central Processing Facilities，简称 CPF）处理后的数据传输到科学操作中心（Science Operations Centre，简称 SOC），SKA1-mid 的所有台站直接通过 CPF 传输到 SOC，后续经由科学处理中心（Science Processing Centre，简称 SPC）分发到分布于各大洲的 SRC，由 SRC 将数据开放授权给科学家和用户。

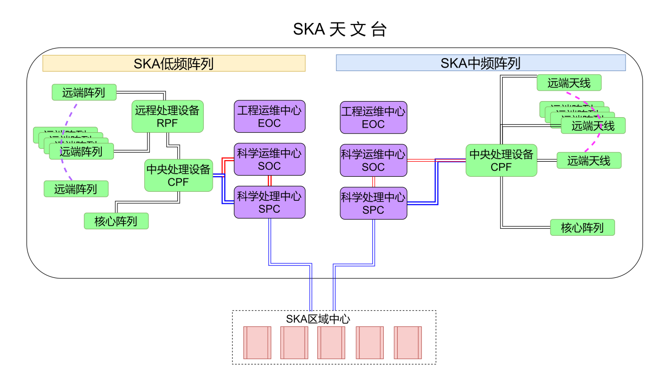


图 1　SKA 观测数据流示意图

Figure 1　Data Flow Diagram of SKA Observation

根据当前最新的设计基线[18]，SKA1 的详细

参数如表1所示，其中的 N_f 为观测频率通道最大值，实际观测中不会超过该值，对于SKA1-mid 观测而言，大多数的观测模式下该值较小，以 SKA-VLBI 为例，频率通道一般 64K 就可以满足观测需求[19]。

表 1　SKA1 阵列参数

Table 1　Parameters of SKA1 Array

	SKA1-low	SKA1-mid
天线数量 N_{ant}	131072	197
积分时间 t_{dump}	0.6s	0.08s
频率通道 N_f	256,000	256,000
波束 N_{beam}	1	1
阵列/口径大小 D_s	35 米	13.5 米/15 米
最大基线 B_{max}	80 公里	150 公里
观测带宽 Δ_v	300MHz	770MHz
分辨率	0.25arcsec	7arcsec

当前 SKA 确定了 5 个首要的科学目标（Key Science Project，简称 KSP），包括宇宙黎明和黑暗时期探测，星系演化、宇宙学与暗能量研究，利用脉冲星和黑洞进行引力的强场检验，宇宙磁场的起源和演化，孕育生命的摇篮，这些 KSP 均包括基础物理学、天体物理学或者宇宙学中尚未解决的问题[1,20,21]，同时也孕育着包括宇宙第一缕曙光的重大科学发现。随着 SKA1 的开工建设，不需要整个 SKA 阵列灵敏度、分辨率或者频率覆盖范围的重要科学主题被确定为高优先级（High Priority Science Objective，HPSO）项目，经过广泛的论证和讨论，我国也确定了 “2+1” 的战略部署，即确保两个优先突破领域和若干具有特色的研究方向，并于 2020 年启动了 2 项 SKA 科学专项，分别为宇宙再电离（the Epoch of Reionisation，简称 EoR）项目和脉冲星项目，EoR 将利用当前 SKA 先导阵列的观测数据再现宇宙的黎明，同时参加大天区的低频巡天，从统计上揭示宇宙再电离时代的宇宙整体结构特性；脉冲星项目主要包括脉冲星搜索（Pulsar Search，简称 PSS）和脉冲星计时（Pulsar Timing，简称 PST），两者主要聚焦于检验引力波理论和探测发现超大质量黑洞的合并事件[1]。

根据当前 SKA 先导阵列 MWA、ASKAP、LOFAR 等的经验，需要传输归档到 SRC 的数据除了数据立方体数据外，还包括整个流水线处理优化的 UV 数据、用于校准流水线的中间产品、UV 格点化数据、点源扩散函数（Point Spread Function，简称 PSF）立方体以及多种分辨率的加权和权重信息等[22–25]。

接下来数据流的预估将主要依据 HPSO 科学目标[10,26]及其观测时间占比，整个流程包括观测数据的初步生成、经过相关处理、预处理等流程。当然除此以外，最终的科学数据还包括项目负责人（Principal Investigator，简称 PI）领导的各个重要观测、其他科学发现以及 SKA-VLBI 项目等。

2.1　观测台站端

阵列中的天线接收到原始的模拟信号后，首先将信号数字化，为了提高信号的时间和频率分辨率，还会使用有限脉冲滤波 (Finite Impulse Response，简称 FIR) 和快速傅里叶变换 (Fast Fourier Transform，简称 FFT) 操作。对于SKA1-low还会进行波束合成来进行信号的校准和时延补偿操作。

此处假定天线数目为 N_{ant}，每个天线有 N_{pol} 路极化输出，信号的采集带宽为 $\Delta\nu$，由奈奎斯特采样定理可知，采样率至少为 $2\Delta\nu$，在使用 N 阶 FIR 时，计算量大约为 $2N_{tap}$，根据 FFT 的算法复杂度 $O(nlog_n)$ 可知，FFT 运算的运算数为 $Nlog_N$，此时每个站经过多项滤波器后需要的操作总数量级为：

$$2N_{tap}N_{ant}N_{pol}\Delta\nu + 2N_{bands}log_2(2N_{band})N_{ant}N_{pol}\Delta\nu_{band}$$

对于波束合成而言，通过给每个天线提供正确的延迟来进行校准，主要用于高时间分辨率和高频率分辨率的科学应用，比如脉冲星、行星际闪烁等，SKA1-low 的相关处理机波束合成部件（Correlator BeamFormer，简称 CBF）接收 512 个站的数据，并生成斯托克斯的四个参数数据。不过此时的校准相对于 FFT 和 FIR 而言，计算运算量较小，此处评估暂时忽略该项，所以此时对于每个观测台站而言，其输出的整体数据速率大约为：

$$R_{station} \propto N_{beams}N_{bands}N_{pol}\Delta f_{band}S_{sample}$$

由上式可知，采集带宽和采样比特数极大地影响了每个台站或阵列的数据流，根据表1进行估算，可知从台站端到达 CSP 的速率约在 Tbps 量级。中频阵列共计 197（其中 64 面属于当前 MeerKAT）面望远镜，原始数据输出速率约为 2TB/s；低频阵列共计 131072 个天线，原始数据输出速率为 158TB/s。

2.2 CSP

CSP 主要用于将来自台站端的数据进行相关处理操作，在相关之前会进行诸如时延补偿的操作，用于消除由于地球自转引起的条纹旋转；同时进行更精细的信道化，用于提高分辨率，并做带通滤波用于矫正台站多项滤波的影响。根据科学目标的不同，划分的精细化通道不一定相同，此处假定为 N_{chan}。在 CSP 还会进行射频干扰的去除（Fadio Frequency Interference，简称 RFI），由于其计算量较小，此处估算暂时不考虑，所以此时 CSP 输出的数据速率主要为台站的两两相关处理产生的：

$$R_{csp} \propto \frac{N_{beams} N_{chan} N_{stat}^2 N_{pol}^2}{t_{dump}}$$

由于数据输出速率与台站数的通道数和台站数目的平方成正比，从表 1可知，对于输入可见度的数据速率预估为 SKA1-low的数据数据速率为 7.15Tbps, SKA1-mid的数据数据速率为 6.4Tbps.

以SKA1-mid 为例，CSP 将来自天线的输入带宽划分为多个频率通道，然后分别做互相关处理，此时的数据流输入为：133 个 SKA 天线（每个 100Gbps 输出带宽）和 64 个 MeerKAT 天线（每个 40Gbps 输出带宽）接收数据，CSP 的总输入带宽约为 57Tbps，由于成像数据处理与脉冲星处理的流程不同，对于成像而言，采用两级信道化，第一级为 512 个粗信道，第二级为 64000 精细化通道；对于脉冲星搜索而言，在 300MHz 的子频带上提供约 75kHz 的信道化通道，将时间分辨率降低到约 ~ 64μs；以 6.4Tbps 的速度输出数据 [18]。

相关处理将输出可见度数据，最大的分辨率在SKA1-mid 的所有基线上可以达到 0.1s 的积分时间。而相关处理的输出结果主要取决于积分时间和通道数目。基线依赖平均（Baseline Dependent Averaging，简称 BDA）技术可以减少基于射电干涉仪基线分布的可见度数据量，通过对全规模 SKA1-low 的模拟，预计使用 BDA 可以将可见度数据量在不同的时间间隔上减少约 50% ~ 85% [27–29]。

2.3 SDP

SDP 接收来自 CSP 输出的数据进行初步成像预处理，虽然目前 SKA 最终的成像策略尚未确定，不过从当前几个 SKA 探路者项目（包括欧洲的 LOFAR、澳大利亚的 MWA 和 ASKAP）来看，将主要包括校准、成像的迭代主循环（Major）和反卷积的迭代次循环（Minor），其中的 Major/Minor 为主要的计算资源部分。经过数次迭代后，SDP 使用改进的天空模型推导出新的校准参数，并开始新的校准循环以进一步增强模型，最后通过恢复残差图像中的源来创建最终的天空图像。SKA1 阶段 SDP 从相关处理设备获取的数据主要包括图像和校准数据等，通常不会保留可见度数据，但对于SKA1-low 的高优先级项目宇宙再电离 EoR 例外，因为该项目需要保留可见度数据才能有效地去除前景干扰。SKA1-mid 的 SDP 接收来自 CSP 的输入数据以及来自望远镜管理（Telescope Manager, 简称 TM）的数据 [30]，并生成包含科学产品和校准产品的数据，主要为连续谱、谱线立方体数据成像数据，脉冲星搜索的候选数据，暂现源探测数据，单天线强度数据以及消除大气和电离层影响的校准数据等。

2.4 SRC **计算负载需求**

SKA 一个重要的挑战就是需要管理和处理来自不同数据处理流程的数据，并产生高动态范围的图像和其他数据产品，提供对后续科学的支持 [31]。根据 SDP 的当前设计，标准的数据产品将根据每个观测项目来制定。对于成像观测而言，均可以在三个维度（两个空间，一个光谱）、偏振及图像中产生最大分辨率的数据产品 [32]，而其中数据量缩减最多的阶段就在将可见度数据进行成像处理的操作过程。

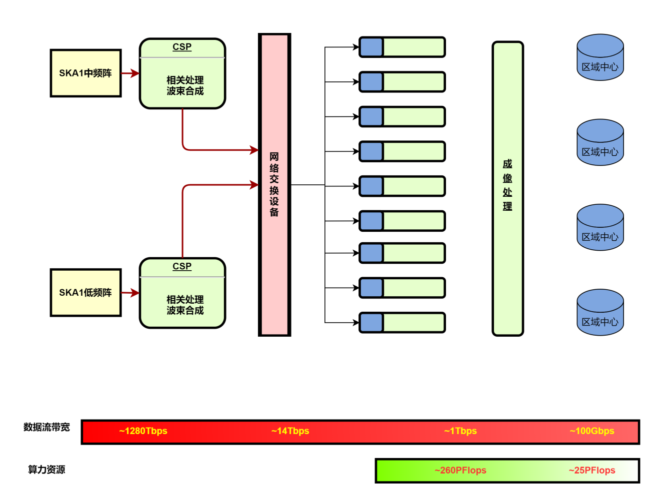


图 2　数据流示意图

Figure 2　Data Flow Diagram

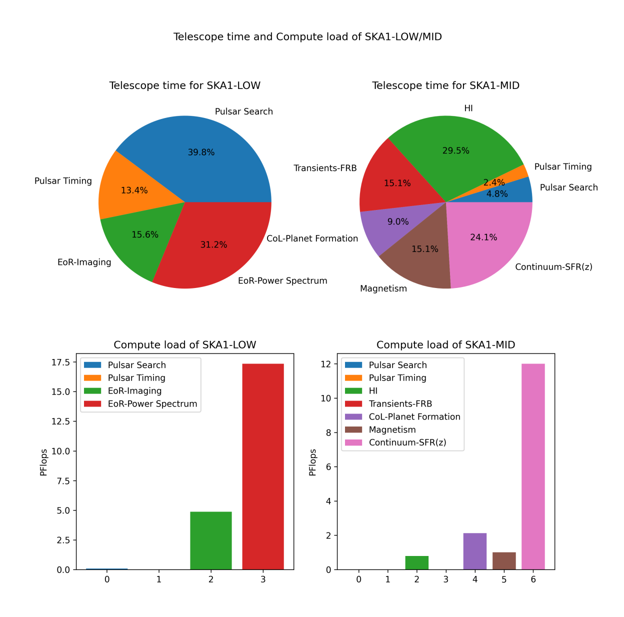


图 3　SKA1 高优先级项目观测时间与计算负载评估

Figure 3　Estimation observate time and compute load of SKA1 HPSO

详细的观测数据流示意图如图1所示，各个阶段的产品数据流示意图如图2所示。台站产生的原始观测数据，在 SDP 经过初步处理生成初步的科学数据产品，主要包括成像的数据（比如图像或者格点化的可见度数据）和其他数据，其中成像数据的大小十分依赖于通道数量，最后的文件大小正比于通道数量，这些数据通过洲际网络分发到每个区域数据中心的子网络，天文学家通过互联网连接到区域数据中心进行数据处理，此时的数据称为高级数据产品（Advanced Data Products，简称 ADP），此时的 ADP 可以通过 SKA 区域中心或者云的方案提供给最终的科研人员。

在SKA-1的所有处理流水线中，成像管道占据了主要的计算需求和资源，如图3 约占SKA1-low 观测时间 15% 和 31% 的 EoR 成像和功率谱观测，对应的计算负载约为 31% 和 55%；非成像处理管道对计算的需求相对较小，比如脉冲星搜索流水线系统（Pulsar Search Pipeline，简称 PSP），约占SKA1-low上 HPSO 时间的 39% 和 SKA1-Mid 的 5% 的时间。对于SKA1-low而言，如果不包含 EoR 的 UV 可见度数据，科学数据速率约为 3Gbps，如果增加 UV 可见度数据，那么数据率将增加到 25Gbps；对于SKA1-mid而言，科学数据速率约为 9Gbps [33]。

以 PSP 为例，仅需要实时校准（Real-time calibration，RCAL）；但对于连续谱成像而言，还需要迭代自校准（Iterative self-calibation，ICAL），生成泰勒项图像的数据准备管线（Data preparation pipeline producing Taylor-term images，DPrepA），生成粗信道化图像的数据准备管线（Data preparation pipeline producing coarse channelised images，DPrepB）；而谱线成像管线还需要额外的生成精细信道化图像的数据准备管线（Data preparation pipeline producing fine channelised images，DPrepC）。

从图3可知，通过相应的加权计算，最后的计算负载约为SKA1-low需要至少 13.6PFLOPS，SKA1-mid需要至少 11.5PFLOPS，所以 SKA1 阶段 SRC 需要的算力预估为 25.1PFLOPS.

3　SKA1 阵列布局

为了实现 SKA 的科学目标，在建设成本与技术方面要均衡考虑，需要建造一个满足极端性能要求而且造价相对可控的望远镜。这需要对当前的科

学需求及各种技术参数（阵列的基线长度、尺寸大小、天线灵敏度、功耗等）进行评估。最终确认了由包含澳大利亚和南非为台址的方案。

在 SKA 的阵列布局和设计中，优先考虑了 KSP 科学观测的需要，通过基线长度的增加，来提高空间分辨率，采用核心阵列，用于增加 UV 覆盖，提高灵敏度。另外在设计中还考虑了包括但不限于频率范围、灵敏度、观测带宽、极化能力、采集面积、基线长度、处理能力等因素。其中最重要的几个因素如下：

- 分辨率：实现高优先级科学所需要的分辨率需求。分辨率取决于频率设置，高频相对于低频而言，更容易实现高空间分辨率。比如 150KM 的阵列，在 8GHz 观测时，其分辨率可以达到 0.0629arcsec
- UV 快照覆盖：基线向量的二维分布决定了瞬时空间频率的覆盖范围，应该尽可能多地覆盖到 UV 平面，在螺旋阵列的布局中，快照覆盖主要由螺旋的旋转角度、螺旋臂上天线的位置以及螺旋臂的数量决定。在图4,5可以看到螺旋的角度以及天线的位置都经过了调整，以提供良好的 UV 覆盖
- 核心区域：核心区域主要考虑提供足够的灵敏度，用于支持脉冲星搜索观测和 HI 线的观测

3.1 SKA1-mid

SKA1-mid望远镜的运行频段为 0.35 ~ 15.3 GHz，位于南非的卡鲁沙漠，主要进行脉冲星、21 厘米中性氢和连续谱的高灵敏度观测等科学研究[18]。该望远镜由 133 个直径为 15 米的偏置格里高利天线，和 MeerKAT 望远镜建造的 64 个直径为 13.5 米的天线组成。SKA1-mid有一个直径为 1 公里的核心阵，一个随机放置的 3 公里的二维阵列和三个螺旋臂，最长基线为 150 公里[5,34]，具体的布局示意图如图4 所示。

3.2 SKA1-low

SKA1-low 望远镜的运行频段为 50 ~ 350 MHz，部署在澳大利亚西部的默奇森射电天文台，主要进行宇宙再电离、脉冲星、太阳系外星系等科学研究[18]。SKA1-low共有 131072 个对数周期天线，这些天线分为 512 个站，每个站 256 个天线。这些站点中的 296 个将被配置到一个中心核心区域，其余 216 个沿三个旋臂部署，最长基线 65 公里[34]。具体的布局示意图如图5所示。

当前的孔径综合阵列技术在进行校准和成像算法方面已经相对成熟，但对于SKA1-mid的量级数据还需要进行软件和算法的研发，才能支持如此海量的数据，并在减少人工输入的情况下得到更好的动态范围[18]。

4 实验测试

本章节列举了使用 OSKAR[2)]进行 SKA1-low 和 SKA1-mid 的模拟数据生成，主要阐述了模拟过程的思路、方法、运行及计算的时间和最后产生的数据文件大小等。

OSKAR 为剑桥大学开发的用于 SKA 模拟的一套软件，基于射电干涉测量方程（Radio Interferometer Measurement Equation，简称 RIME）模拟生成可见度数据，主要功能为模拟台站波束、干涉阵列数据，可以生成对应的二进制可见度数据或测量集数据（Measurement Set，简称 MS），同时还可以将 FITS 文件转变为天空模型、对可见度数据增加噪声等功能。

模拟实验的策略为选取不同的天空模型源数目，分别为 9,100,1024,10000；观测时间从 1 秒到 24 小时，分别为 1s，6s，60s，600s,1200s，2400s，3600s，7200s，14400s，28800s，43200s，54000s，72000s，86400s，频率通道默认为 1，SKA1-mid的观测频率选定为 1.4GHz，带宽为 1KHz，SKA1-low的观测频率选定为 200MHz，带宽为 30MHz。

4.1 SKA1-mid模拟数据测试

对于SKA1-mid的模拟试验，因为计算量不高，故使用了 2 种不同的 GPU 板卡，用于评估整套模拟系统的可扩展性和稳定性，板卡分别为 NVIDIA 的 Tesla K40M 和 Tesla K20X. 其中 Tesla K20X

2) https://github.com/OxfordSKA/OSKAR

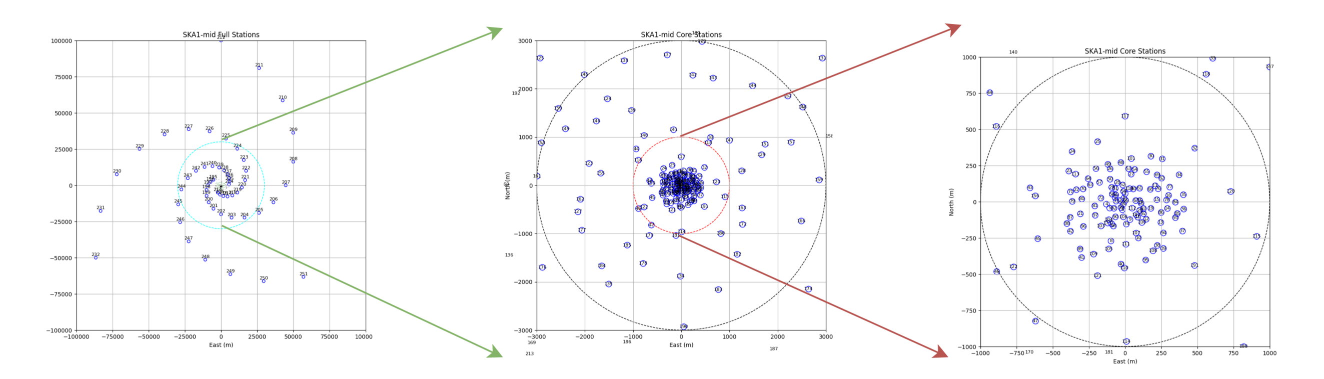


图 4　SKA1 中频阵列布局示意图

Figure 4　Layout diagram of SKA1-mid Array

基于 Kepler 架构，有 2688 个核，具备 6GB 的内存，总线带宽可达 249.6GB/s，双精度的理论值为 1312GFLOPS；Tesla K40m 基于 Kepler 架构，有 2880 个核，具备 12GB 的内存，总线带宽可达 288.4GB/s，双精度的理论值为 1682GFLOPS；可知从具备的核心数、内存、总线带宽和理论值，Tesla K40m 均优于 K20X，从对阵列的实测结果图6,7,8也可以看到，K40m 的试验参数比 K20X 有所提升，不过提升不大，主要因为两个板卡属于同一个系列，不过 K40m 还是具有优势的，在多点模拟过程中，消耗的时间均低于对应的 K20X 显卡。

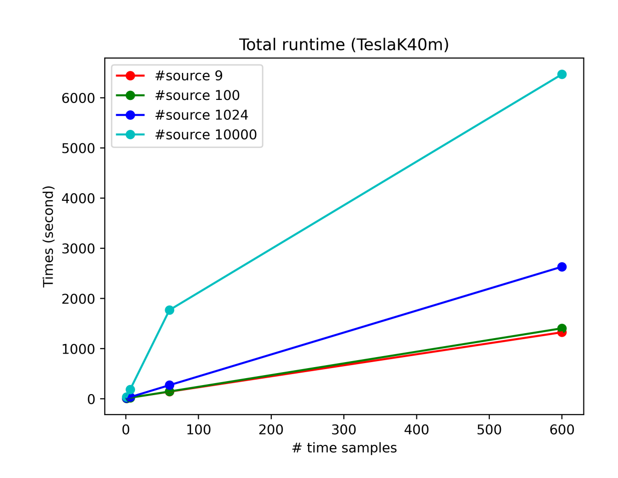


图 7　SKA1 中频阵列 K40m 模拟

Figure 7　Simulation of SKA1-mid based on Tesla K40m

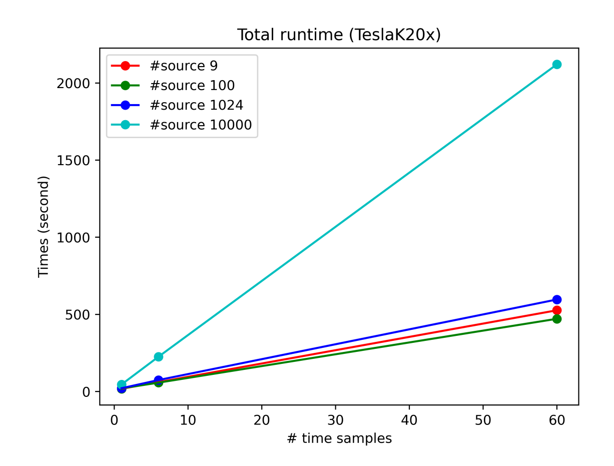


图 6　SKA1 中频阵列 K20X 模拟

Figure 6　Simulation of SKA1-mid based on Tesla K20X

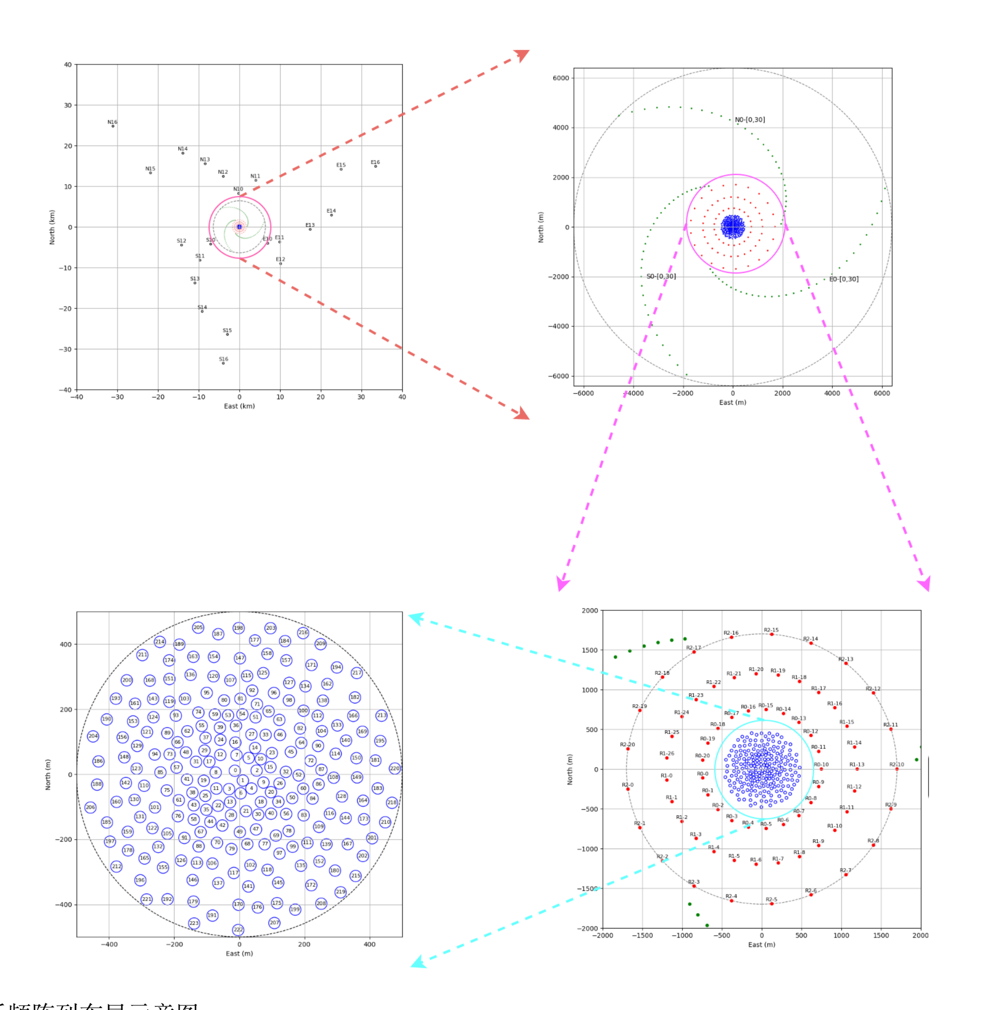


图 5　SKA1 低频阵列布局示意图

Figure 5　Layout diagram of SKA1-low Array

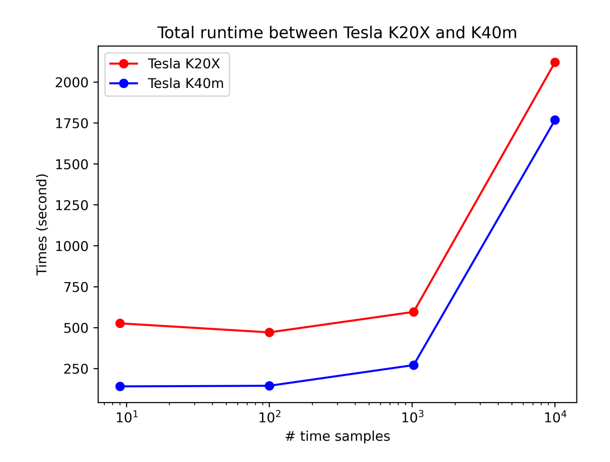


图 8　SKA1 中频阵列 K20X 与 K40m 模拟对比

Figure 8　Simulation compare of SKA1-mid between Tesla K20X and K40m

4.2　SKA1-low**模拟数据测试**

由于SKA1-low的台站较多，算力需求比较大，将使用 China-SRC 平台进行上述模拟测试，CSRC-P搭载的 GPU 型号为 NVIDIA Tesla V100 SXM2，该 GPU 采用 Volta 架构，有 5120 个核，具备 32GB 内存，总线带宽可达 897GB/s，双精度的理论值为 7834TFLOPS。在对SKA1-low进行模拟时，可以看到对于观测时长在 2400s 以内的数据，模拟时间均在 100 秒以内，充分显示了 V100 芯片在计算方面的强大优势。

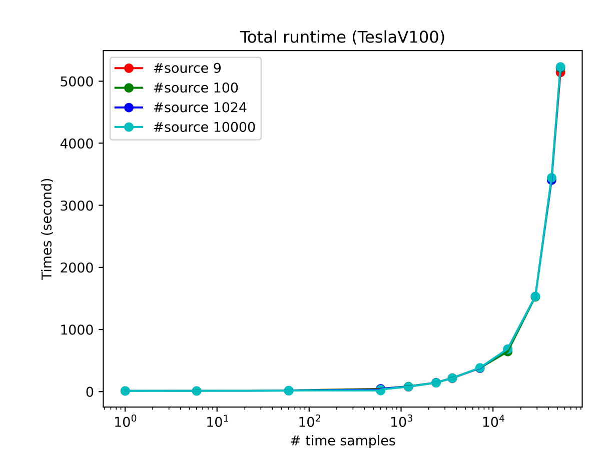


图 9　SKA1 低频阵列模拟数据生成消耗时间
Figure 9　Simulation data generation time consumption of SKA1-low Array

从表2看到,最终的模拟数据大小与天空模型中的源相关性不大,模拟所花费的时间也基本一致,模拟的时间与最终生成的可见度数据的大小，主要取决于观测的时间。并且通过设定观测的频率通道可知，文件的最终大小正比于该值，对于SKA1-low模拟而言，如果使用表1中最大的频率通道，每秒产生的可见度数据大小约为 2.048GB，约 2TB/s。

5　总结与展望

SKA 无疑将是有史以来最大的射电观测基础设施之一，通过对 SKA1 整套系统的数据流进行了基于模型的预估、分析和模拟，其 Tbps 的数据流、PB 级的存储归档需求及观测复杂性，对后续的数据管理、处理、计算、存储和网络都提出了最严峻的挑战，通过对整个数据流的讨论，为后续的中国 SKA 区域中心的正式运行提供数据系统级的支持。项目将结合射电天文技术以及包括信息、通信、计算机等现代信息技术。SKA 的超大数据流将极大地改变天文研究的方式，而使 SKA 成功运行则需要对传统框架设计进行革命性的改进和突破。在 SKA1 阶段就需要联合电子学、计算机技术、信息等领域多个学科的专业知识来进行应对。SKA 也将特别推动信息技术、天文软件和通讯技术等的发展。

致谢　向评审人和对该文有帮助的人士表示谢意. 本研究使用了由国家重点研发计划大科学装置前沿研究专项 (2018YFA0404603) 资助研制的中国 SKA 区域中心原型机的资源。

参考文献

1　武向平主编. 中国 SKA 白皮书 (中文版). 北京：科学出版社，2017
2　中国参加 SKA (第一阶段) 综合论证报告. 技术报告，科学技术部. 2018
3　An T, Wu X P, Hong X Y, SKA data take centre stage in China. Nat Astron, 2019, 3: 1030–1030
4　Guo S G, Zheng X Y, Mao Y F, et al. Scheme and Prospect of the SKA Big Data Transferring(in Chinese). E-science Technology & Application, 2018，9(3): 3–13 [郭绍光，郑小盈，毛羽丰，等. SKA 海量数据传输的方案及展望. 科研信息化技术与应用，2018，9(3): 3–13]
5　Swart G P, Dewdney P E. Highlights of the SKA1-Mid Telescope architecture. Journal of Astronomical Telescopes, Instruments, and Systems, 2022, 8(1): 011021
6　Quinn P, van Haarlem M, An T, et al. SKA Regional Centres White Paper v1.0. Technical Report, Square Kilometre Array. 2020
7　Guo S G, An T, Xu Z J, et al. Progress and Prospect of transcontinental high-speed data transmission at SKA Regional Center in China(in Chinese). ChinaXiv:T202206.00291 (2022) [郭绍光, 安涛，徐志骏，等. 中国 SKA 区域中心跨洲际高速数据传输进展及展望. ChinaXiv:T202206.00291 (2022)].
8　Jongerius R, Wijnholds S, Nijboer R, et al. An End To End Computing Model for the Square Kilometre Array. Computer, 2014, 47(9): 48–54
9　An T, Wu X C, Lao B Q, et al. Status and progress of China SKA Regional Centre prototype. arxiv:2206.13022

表 2 SKA1 低频阵列模拟数据大小 (字节)

Table 2 SKA1-low simulation filesize (Bytes)

观测时间丨天空源数目	9	100	1024	10000
1	8442505	8442023	8442025	8442024
6	50365065	50364583	50364584	50364584
60	503178874	503178392	503178394	503178394
600	5030775854	5030775370	5030775370	5030775372
1200	10061493857	10061493377	10061493373	10061493375
2400	20122929856	20122929375	20122929381	20122929377
3600	30184365864	30184365380	30184365381	30184365385
7200	60368673866	60368673380	60368673379	60368673381
14400	120737289864	120737289385	120737289380	120737289388
28800	241474521875	241474521391	241474521386	241474521386
43200	362211753879	362211753392	362211753398	362211753392
54000	452764677879	452764677396	452764677394	452764677392
72000	603686217875	603686217395	603686217390	603686217398
86400	724423449873	724423449399	724423449400	724423449397

10 Quinn P, Axelrod T, Bird I, et al. Delivering SKA science. In: Proceedings of Advancing Astrophysics with the Square Kilometre Array (AASKA14). Giardini Naxos, 2015

11 An T. Science opportunities and challenges associated with SKA big data. Science China: Physics, Mechanics & Astronomy, 2019, 62(8): 121–126

12 Hollitt C，Johnston-Hollitt M，Dehghan S, et al. An Overview of the SKA Science Analysis Pipeline. In: Astronomical Data Analysis Software and Systems XXV. San Francisco: Astronomical Society of the Pacific, 2017. 367–370

13 Lao B Q, Zhang Y K, An T, et al. Software Platform on China SKA Regional Center Prototype System(in Chinese).ChinaXiv:202206.00173. [劳保强，张迎康，安涛，等.(2022). 中国 SKA 区域中心原型系统 – 软件平台.ChinaXiv:202206.00173]

14 Xu Z J, An T, Guo S G, et al. A machine learning dataset for FRB detection in raw data(in Chinese).ChinaXiv:T202206.00321.[徐志骏，安涛，郭绍光，等.(2022). 一个面向原始数据搜寻的快速射电暴数据集.ChinaXiv:T202206.00321]

15 Wei J W, Zhang C F, Zhang Z L, et al. Parallel optimization of the pulsar search pipeline on China SKA Regional Centre Prototype (in Chinese). ChinaXiv:T202206.00297. [韦建文, 张晨飞，张仲莉, 等.(2022). 低频射电脉冲星搜索的性能优化方法.ChinaXiv:T202206.00297]

16 Wei J W, Zhang C F, Lao B Q, et al. Optimization of parallel processing of Square Kilometre Array low frequency imaging pipeline (in Chinese). ChinaXiv:T202206.00292. [韦建文, 张晨飞, 劳保强, 等.(2022).SKA 低频成像管线并行优化化.ChinaXiv:T202206.00292]

17 Braun R, Bourke T L, Green J A, et al. Advancing astrophysics with the square kilometre array. In: Proceedings of Advancing Astrophysics with the Square Kilometre Array (AASKA14). Giardini Naxos, 2015. id.174

18 Dewdney P. SKA1 System Baseline Design. Technical Report, SKA Organisation. 2016

19 Operational Model for inclusion of SKA in Global VLBI,JIV-ERIC and Square Kilometre Array Phase 1 Project, Submission date: 15/6/2020

20 Carrilli C L, Rawlings S. Science with the Square Kilometer Array: Motivation, Key Science Projects, Standards and Assumptions. New Astronomy Reviews, 2004, 48(11): 979–984

21 Bourke, T, Braun, R, Fender, R, et al. (2015). Advancing Astrophysics with the Square Kilometre Array (AASKA14).

22 Whitney A, Booler T, Bowman J, et al. The Murchison Widefield Array (MWA): Current Status and Plans. In: American Astronomical Society, AAS Meeting #218. Bulletin of the American Astronomical Society, 2011. Vol. 43, id.132.07

23 Haarlem M P V, Wise M W, Gunst A W, et al. LOFAR: The Low Frequency Array. Astronomy & Astrophysics, 2013, 556(7):

629–635

24 Jonas J L. MeerKAT as an SKA Pathfinder. Jonas J L. The MeerKAT SKA precursor telescope. In: Proceedings of Panoramic Radio Astronomy: Wide-field 1-2 GHz research on galaxy evolution. Edited by G. Heald and P. Serra. Groningen, 2009. id.4

25 Johnston S, Bailes M, Bartel N, et al. Science with the Australian Square Kilometre Array Pathfinder. Publications of the Astronomical Society of Australia, 2007, 24(4): 174–188

26 Ford D, Bolton R C, Colegate T, et al. The SKA Cost/Performance Tool: A Hierarchical SKA Modelling Tool. In: Proceedings of Wide Field Astronomy & Technology for the Square Kilometre Array. Chateau de Limelette, 2009. id.22

27 Cotton W D. Special problems in imaging. In: Synthesis Imaging, ed. R. A. Perley, F. R. Schwab, & A. H. Bridle. 1986. 123–136

28 Cotton W D. Special Problems in Imaging. In Astronomical Society of the Pacific Conference Series, Vol. 180, Synthesis Imaging in Radio Astronomy II, ed. G. B. Taylor, C. L. Carilli, & R. A. Perley. 1999. 357–370

29 Deng Q W, Wang F, Deng H, et al. Performance evaluation of baseline-dependent averaging based on full-scale SKA1-LOW simulation. Research in Astronomy and Astrophysics, 2022, 22(4): 045014

30 Natarajan S, Barbosa D, Barraca J P, et al. SKA Telescope Manager (TM): status and architecture overview. In: Proceedins of the SPIE. 2016. 9913: id.991302

31 Torchinsky S A, van Ardenne A, van den Brink-Havinga, et al. SKA Data Flow and Processing. ?Alexander P, Bregman J A, Faulkner A J. SKA Data Flow and Processing. In: Proceedings of Wide Field Astronomy & Technology for the Square Kilometre Array. Chateau de Limelette, 2009. id.16

32 Spekkens K, Chiang C, Kothes R, et al. Final Report to LRP Panel: the Square Kilometre Array.

33 Broekema P C, van Nieuwpoort R V, Bal H E. The square kilometre array science data processor. Preliminary compute platform design. Journal of Instrumentation, 2015, 10(7): C07004.

34 Grainge K, Alachkar B, Amy S, et al. Square Kilometre Array: The radio telescope of the XXI century. Astron Rep, 2017, 61, 288–296

Scientific data flow and array simulation analysis for the SKA-1 era

GUO ShaoGuang[1,2*], Lu Yang[1], AN Tao[1,2], LAO BaoQiang[1,2], XU ZhiJun[1,2], WU Xiaocong[1,2] & LV WeiJia[1,2]

1. SKA Regional Centre Joint Lab, Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Shanghai 200030, China;
2. SKA Regional Centre Joint Lab, Peng Cheng Laboratory, Shenzhen, 518066, China

After years of planning for the next generation of radio telescopes, the Square Kilometer Array (SKA), the construction of the SKA phase one (SKA1) had started in July 2021. After the formal operation of SKA1, it is expected that 750 petabytes of scientifically processed data will be generated every year. The data will be stored at SKA regional centers around the world for further analysis by researchers. In this paper, the models of SKA observation station, central signal processor, scientific data processing and regional center are quantitatively analyzed. Based on the high-priority scientific observation of SKA1, the data flow evaluation at each stage and the demand for computing power of scientific data processing are obtained. Taking the current SKA1-low and SKA1-mid arrays as examples, the key factors affecting the layout of interference arrays including resolution, sensitivity and UV coverage are summarized. Finally, OSKAR is used for data simulation of interference array. Through the simulation of SKA1-mid, the scalability and stability of the system are obtained. Through the simulation of SKA1-low on CSRC-P, it can be seen that the design of prototype SKA regional center in China has been fully optimized. And the detailed requirements of computing power and the detailed information of data volume are obtained. The SKA's demand for data processing, computing and storage also requires a combination of technologies and interdisciplinary efforts from areas such as electronics, communication, information technology and computer.

Sqare Kilometre Array , Data simulation, Synthesis array, Data format

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

doi: ??